IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Addressing the overlapping data problem in classification using the One-vs-One decomposition strategy

JOSÉ A. SÁEZ[1], MIKEL GALAR[2], AND BARTOSZ KRAWCZYK[3].

[1]Dept. of Computer Science and Automatics, University of Salamanca, Plaza de los Caídos s/n, 37008 Salamanca, Spain (e-mail: joseasaezm@usal.es)
[2]Institute of Smart Cities (ISC), Public University of Navarre, 31006, Pamplona, Spain (e-mail: mikel.galar@unavarra.es)
[3]Dept. of Computer Science, Virginia Commonwealth University, 401 West Main Street, Richmond, VA 23284, USA (e-mail: bkrawczyk@vcu.edu)

Corresponding author: José A. Sáez (e-mail: joseasaezm@usal.es).

ABSTRACT Learning good-performing classifiers from data with easily separable classes is not usually a difficult task for most of algorithms. However, problems affecting classifier performance may arise when samples from different classes share similar characteristics or are *overlapped*, since the boundaries of each class may not be clearly defined. In order to address this problem, the majority of existing works in the literature propose to either adapt well-known algorithms to reduce the negative impact of overlapping or modify the original data by introducing/removing features which decrease the overlapping region. However, these approaches may present some drawbacks: the changes in specific algorithms may not be useful for other methods and modifying the original data can produce variable results depending on data characteristics and the technique used later. An unexplored and interesting research line to deal with the overlapping phenomenon consists of decomposing the problem into several binary subproblems to reduce its complexity, diminishing the negative effects of overlapping. Based on this novel idea in the field of overlapping data, this research proposes the usage of the *One-vs-One* (OVO) strategy to alleviate the presence of overlapping, without modifying existing algorithms or data conformations as suggested by previous works. To test the suitability of the OVO approach with overlapping data, and due to the lack of proposals in the specialized literature, this research also introduces a novel scheme to artificially induce overlapping in real-world datasets, which enables us to simulate different types and levels of overlapping among the classes. The results obtained show that the methods using OVO achieve better performances when considering data with overlapped classes than those dealing with all classes at the same time.

INDEX TERMS Classification, data generation, decomposition strategies, one-vs-one, overlapping data.

I. INTRODUCTION

In a *classification* problem a series of input attributes must be linked to a discrete output class [18], [44]. This relationship is established by learning classifiers, which are models built from a set of labeled samples of the problem. It is usually not a problem to obtain good-performing classifiers when classes are easily separable. However, in real-world data samples from different classes may share similar attribute values [33]. In these cases, the boundaries of the classes may not be clearly defined, being too complex to be correctly learned. This problem is commonly referred as *overlapping data* [16], [40]. These samples cause uncertainty when determining the decision boundaries and thus, negatively affect classification performance [16].

The existing proposals in the specialized literature to overcome this problem are based on two main strategies:

1) *Adaptation of classification algorithms*. Some works adapt well-known methods to mitigate the effects produced by overlapping in classification. Thus, for example, Fu et al. [11] and Czarnecki and Tabor [7] propose to adapt *Support Vector Machines* (SVM) [3] to deal with overlapping data, whereas Xiong et al. [40] focus on modifications of the *Naïve Bayes* [21] algorithm.
2) *Data preprocessing*. These works alter the original data aiming to reduce the impact of overlapping in classifier performance [23]. The original data is modified either separating the overlapping classes by the introduction of complementary features or merging overlapping

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2925300, IEEE Access

classes to form meta-classes.

Even though both approaches can improve classifier performance in specific scenarios, they present some drawbacks. The former is based on modifying an existing method, which may be sometimes hard to perform. Moreover, since the improvement comes from the adaptation of the method, it is not directly applicable to other algorithms [11]. Otherwise, the latter involves the usage of preprocessing techniques, which are time-consuming and are usually designed to deal with data having particular characteristics [23]. Hence, to avoid these shortcomings, other approaches to reduce the impact of overlapping need to be studied, which neither involve algorithm modifications nor making assumptions about data characteristics.

When working with multi-class problems, the usage of *binary decomposition strategies* [24] has not been yet considered as an alternative to deal with overlapping data. However, it may be an interesting alternative to the aforementioned approaches. Decomposition strategies divide the original problem into several two-class subproblems as a way to reduce the original complexity [17], [46]. Among these strategies, the *One-vs-One* (OVO) decomposition [45], [46], which divides the original problem into as many subproblems as possible pairs of classes, is one of the most widely used schemes in the literature [13], [28]. This research analyzes the suitability of OVO to deal with overlapping data. Since only two classes are considered in each subproblem, OVO will be able to increase the separability between them, reducing the impacts of overlapping and thus, improving the final classification performance.

However, in order to properly evaluate the benefit of using OVO to deal with overlapping we come to a bigger problem, the lack of evaluation frameworks for overlapping data problem. For this reason, in this paper we also provide a new and systematic way to introduce overlapping into real-world datasets so that methods dealing with overlapping can be properly evaluated. This new framework is introduced in Section IV and will allow as to fairly evaluate the difference between the usage of OVO and not using it. With this framework one can control the amount of overlapping in the data and one can exactly determine which samples in the data belong to the borderline, the overlapping and non-overlapping regions. This rigorous identification of the different types of samples in the dataset implies a completely novel way of understanding and evaluating the overlapping data problem in the literature. This way, we will be able to perform a thorough analysis and extract conclusions on classifier performance in each region (see Sections VI-VII).

The suitability of the OVO decomposition with overlapping data is analyzed in an extensive empirical study considering well-known learning algorithms, such as C4.5 [27], *Repeated Incremental Pruning to Produce Error Reduction* (RIPPER) [6], *k-Nearest Neighbors* (k−NN) [5] and SVM [3]. We will analyze the differences between applying standard and OVO-based classifiers over a total of 1394 datasets with different degrees of overlapping. The different regions will be considered to analyze the effect of overlapping in the classifiers' performance. The robustness of the methods in terms of its *Equalized Loss of Accuracy* (ELA) metric [30] will be also studied. In total, more than 2,091,000 results will be analyzed and will serve as a solid basis to establish a comparison between the OVO and non-OVO versions of the classifiers. The main lessons learned in this research, including interesting findings related to the experimentation performed and its analysis are summarized in Section VIII. A web-page with the datasets and the results obtained for each classification algorithm is available at https://joseasaezm.github.io/overlapping/.

The rest of this work is organized as follows. Section II introduces decomposition strategies and the OVO model as a possible solution for overlapping data. Section III presents related works on overlapping data in classification. Section IV describes the proposed scheme for introducing overlapping in real-world data. Then, Section V presents the experimental framework. Section VI analyzes the results obtained when overlapping data affects training and test sets, whereas Section VII focuses on results with overlapping only in training sets. Section VIII summarizes the main findings of our empirical study. Finally, Section IX presents the concluding remarks.

II. BINARY DECOMPOSITION STRATEGIES FOR DATASETS WITH MULTIPLE CLASSES

Multi-class data [1], [42] are frequent in real-world tasks, being a generalization of data with only two classes (*binary* problems). Multi-class classification data have been traditionally addressed following two different approaches [24]:

1) *Algorithm level approaches*. They adapt methods that learn from binary data to deal with more classes [12].
2) *Data decomposition approaches.* They decompose multi-class problems into binary subproblems, reducing the complexity of the original problem [13].

Modifying existing methods to deal with multi-class data may be a complex task in some cases, e.g. when working with SVM [3]. Data decomposition can be used in such scenarios, since any binary classification algorithm can be employed as base learner without adapting its learning procedure. In this section, we first introduce decomposition strategies and its advantages [13] (Section II-A) and then focus on the OVO decomposition (Section II-B).

A. DECOMPOSITION OF MULTI-CLASS PROBLEMS

Using binary decomposition in multi-class problems usually carries certain benefits [13], [28]. First, they enable algorithms designed to deal with binary data to address multi-class problems [24]. Another advantage, which this research takes advantage of when dealing with overlapping data, is that the separation of the different classes becomes less complex using decomposition [12]. Thus, decomposition allows classes in certain classification problems to be more easily separable when considered in pairs [2], [17], [45].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2925300, IEEE Access

On the other hand, decomposition strategies lead to the formation of ensembles of classifiers, which are considered as one of the most powerful techniques in contemporary machine learning [38].

Binary decomposition is based on two main phases [13]:

1) *Problem division* [24]. The data are split into binary subproblems that are then treated by binary classifiers [12]. Two main decomposition strategies exist [24]:
   - *One-vs-One* (OVO) [45], [46] splits a problem with C classes into $C(C-1)/2$ subproblems, training a different classifier for each pair of classes.
   - *One-vs-All* (OVA) [17], [31] splits a problem with C classes into C subproblems, training a different classifier to distinguish each class from the others.
2) *Output combination* [13]. To classify new samples, they are presented to all the classifiers and their outputs are combined to obtain the final result. Among the combination methods found in the literature, *Weighted Voting* [20], *probability estimates* [39] and *majority voting* [13] should be highlighted.

This research focuses on OVO due to its proven advantages with respect to OVA [28], such as the creation of simpler borders between classes, the increase in classification performance and the shorter training times when working with smaller subproblems. Finally, OVA may also create imbalanced datasets, which is known to be a major problem in machine learning [13].

B. ONE-VS-ONE BINARY DECOMPOSITION

OVO splits a dataset with C classes into $C(C-1)/2$ binary problems. Each binary problem consists of those training samples involving the pair of classes (c_i, c_j) with $i < j$. Then, a classifier is built for each one of these binary problems.

New samples are classified by being submitted to all classifiers. A classifier, distinguishing between c_i and c_j, computes a confidence $r_{ij} \in [0,1]$ in favor of c_i (r_{ji} is computed as $1 - r_{ij}$). These confidences are stored in a *score matrix*:

$$R = \begin{pmatrix} - & r_{12} & \cdots & r_{1C} \\ r_{21} & - & \cdots & r_{2C} \\ \vdots & \vdots & \vdots & \vdots \\ r_{C1} & r_{C2} & \cdots & - \end{pmatrix} \quad (1)$$

Finally, combination methods [13] are employed to compute the class label of new samples from the score matrix. Among them, majority voting is used in this work. It is one of the most used and simplest approaches, based on predicting the class with the largest number of votes by the classifiers. This approach has shown to provide a similar behavior to more complex strategies [13]. Using the majority voting scheme, the final class label can be computed as:

$$Class = \arg\max_{i=1,\ldots,C} \sum_{1 \le j \ne i \le C} s_{ij}, \quad (2)$$

where $s_{ij} = 1$ if $r_{ij} > r_{ji}$ and $s_{ij} = 0$ otherwise.

III. CLASS OVERLAPPING AFFECTING CLASSIFICATION PROBLEMS: RELATED WORKS

Real-world data usually involve overlapping among samples of different classes [4], [40]. This fact implies that some samples of a class c_i have similar characteristics to those of a different class c_j. The area of the domain in which these specific samples are found is called the *overlapping region* [22]. All the samples belonging to this region are characterized by having non-zero probability densities for each class.

Some works have shown that many classification errors occur in the boundaries of the classes, which may be altered by the presence of overlapping samples [33], [40]. This fact may increase the chances of incorrect predictions [16]. Given the loss of accuracy to which these types of samples can lead, methods that can alleviate class overlapping are of special interest [11], [40].

As it was mentioned, some proposals adapt classification algorithms [7], [11], [40] or modify the original data including additional features [23] to mitigate the impact of overlapping. Other works propose the usage of soft decision strategies assigning multiple class labels to the samples of the overlapping region, which can be then analyzed [32], [34].

A large part of the literature studying the overlapping problem also focuses on imbalance data [19], [35]. Although learning difficulties in class imbalance have been traditionally related to bias towards the majority class, some works show that these are more linked to other factors related to data characteristics such as overlapping [4], [33]. For example, Prati et al. [25] developed a study using a set of artificial datasets showing that the degree of class overlapping has a strong correlation with class imbalance. In this scenario, the use of over-sampling methods based on SMOTE [10], [41] has shown to be very effective [4]. The large influence of overlapping on classification performance with respect to the imbalance ratio was also corroborated in the particular case in which the minority class is more represented in the overlapping region than the majority class [14], [15]. Other proposals to deal with overlapping in imbalanced data include removing the samples belonging to the overlapping region [43] or adapting classification methods [26].

An important aspect of the aforementioned works is the way in which they estimate or control the level of overlapping in real-world datasets. Most of the studies do not take this issue into account, which limits their insight into the nature of the problem. Controlling the level of overlapping in datasets enables the possibility of thoroughly analyzing the properties and robustness of the examined methods. Because of this, some works try to quantify the overlapping level of each real-world dataset considered. For example, some of them compute basic statistics for each attribute [11]. Other works consider more complex metrics such as the *Fisher's discriminant ratio* or the *Kullback-Leibler divergence* between classes [36].

Many works complement their experiments creating synthetic datasets [4], [15], [36]. These types of data have

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2925300, IEEE Access

the advantage of making the level of overlapping easier to control. However, most of the datasets generated in these works have two dimensions and two classes. The basic idea is to create two clusters of samples, one per class, which are initially separated when no overlapping level is considered. Then, an increase of the overlapping level implies that the distance between the cluster centroids is reduced, making them overlap more and more. Clusters with rectangular [14], [15] or circle-like shapes [4], [25] are the most common options in the literature.

The use of each type of data, either real-world or synthetic, offers different advantages:

- *Data variety and complexity in real-world datasets.* Considering real-world data provides a great variety of choice, since each dataset is different and has different properties, which usually imply a greater complexity and richness of characteristics. This cannot be generally achieved by the synthetic data generators proposed in the literature.
- *Overlapping level in synthetic datasets.* Synthetic data allow to control the overlapping level and extract conclusions based on it. Quantifying the level of overlapping is not always easy in real-world data, which only have a specific quantity of overlapping samples, and the effects in classifier performance of varying levels of overlapping cannot be measured.

For these reasons, a systematic way to combine the advantages of both alternatives is required. It would be interesting to have the possibility of introducing, in a supervised manner, different degrees of overlapping in real-world datasets. This fact leads us to our proposal in Section IV, a new scheme for introducing overlapping in real-world data.

IV. A NOVEL SCHEME FOR CREATING OVERLAPPING REGIONS IN REAL-WORLD CLASSIFICATION DATASETS

This section presents a new scheme designed to introduce overlapping in any real-world problem. Section IV-A details the process to create a set of synthetic overlapping samples S for a specific class in the original dataset. Then, Section IV-B describes how the overlapping dataset is built considering S and the original data, giving a mathematical description on how the sets of samples belonging to the overlapping and non-overlapping regions are composed. Finally, Section IV-C presents two possible schemes to introduce overlapping in real-world data depending on which classes are affected.

A. GENERATING THE OVERLAPPING REGION

The overlapping introduction scheme generates an overlapping level $x\%$ affecting one of the classes c of the dataset D. This fact implies that samples from other classes (different from c) invade the domain corresponding to class c, starting from the class boundary of c to its core.

Algorithm 1 shows the pseudocode of the procedure to create the set of synthetic samples S from the original data is D. Notice that the overlapping consists of adding new samples and is not a mere modification of existing samples. This is made this way to avoid altering the underlying class structures of the original data.

The creation of the synthetic samples forming the overlapping region is based on two main steps. They are described afterwards, referring to the associated lines in Algorithm 1:

1. Estimation of the distance of each sample of the target class to the borderline region (lines 1-7). This first step identifies those samples of the target class that are closer to the class boundaries. The closer to the class boundaries are, the more probability to form the overlapping region have (created in the second step).

With this aim, for each sample e_i belonging to class c, the average distance d_i to its k_1 closest samples of the other classes is computed (line 4). Likewise, the majority class m_i of these neighbors is computed (line 5). In this work, we fix a value of $k_1 = 1$ to compute the distance of each sample of class c to the class boundaries –note that higher values of k_1 could be chosen to reduce the negative effect of noisy samples in the dataset. `HVDM` [37] is used to compute the distance between samples.

The distances and classes computed in this step are used in the next step. Our assumption is that the lower the distance d_i associated to each sample e_i is, the closer this sample to the borderline region is. As a result of this step, all the possible triplets $\{(e_i, d_i, m_i)\}$ are computed (line 6).

2. Generation of the synthetic samples to form the overlapping region (lines 8-14). The number of synthetic samples to be introduced (M) is based on the quantity of samples of the target class c, being its $x\%$ (line 8). Then, M samples of class c are sequentially chosen, sorted in a ascending manner by d_i (lines 9-10). For each one of these samples, a synthetic sample s_i is created in its neighborhood (lines 11-12). To this end, a random neighbor n_i among any of its k_2-nearest neighbors is chosen and the sample s_i is created following an interpolation scheme similar to that used by `SMOTE` [10], [41], in which e_i is the sample of the target class c, n_i is the selected neighbor and r is a random number in $(0, 1)$ following an uniform distribution (Figure 1). For nominal attributes, a random value between those of e_i and n_i is chosen.

The value of k_2 ($k_2 = 3$ in our experiments) used to compute the nearest neighbors determines the size of the area around the sample e_i in which the synthetic sample s_i will be created. The value $k_2 = 3$ is used trying to introduce some randomness when creating the new synthetic sample. Considering higher values for k_2, such as $k_2 = 5$ or $k_2 = 7$, may imply the risk of creating the synthetic samples too far from the area of interest in the decision boundaries in which we want to introduce the overlapping data.

B. BUILDING OF THE FINAL OVERLAPPING DATASET

The final dataset with overlapping O will be formed of the samples from D and S ($O = D \cup S$). The overlapping

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2925300, IEEE Access

Input: original dataset D, target class c, overlapping level $x\%$.
Output: set of synthetic overlapping samples S.

1 Set $T = \{e_i | e_i \in D$ and e_i belongs to class $c\}$;
2 **for** *each sample* $e_i \in T$ **do**
3 $\quad N_{e_i} \longleftarrow$ Set of the k_1-nearest neighbors of e_i in D of class $\neq c$;
4 $\quad d_i \longleftarrow$ Average distance from e_i to the samples of N_{e_i};
5 $\quad m_i \longleftarrow$ Majority class of the samples of N_{e_i};
6 $\quad B \longleftarrow B \cup \{(e_i, d_i, m_i)\}$;
7 **end**
8 Compute the number of overlapping samples to be created $M \longleftarrow (\#T \cdot x)/100$;
9 $B' \longleftarrow$ first M triplets $\{(e_i, d_i, m_i)\}$ of B with minimum distance d_i $(i = 1, \ldots, M)$;
10 **for** *each triplet* $\{(e_i, d_i, m_i)\} \in B'$ **do**
11 $\quad n_i \longleftarrow$ Pick out one random sample in D among the k_2-nearest neighbors of e_i;
12 $\quad s_i \longleftarrow$ Compute the synthetic sample with class m_i by interpolation using e_i and n_i;
13 $\quad S \longleftarrow S \cup \{s_i\}$;
14 **end**

Algorithm 1: Proposal to create the set of overlapping samples S (initially $B = S = \emptyset$; k_1 and k_2 are two prefixed parameters indicating the number of neighbors considered in different steps of the proposal).

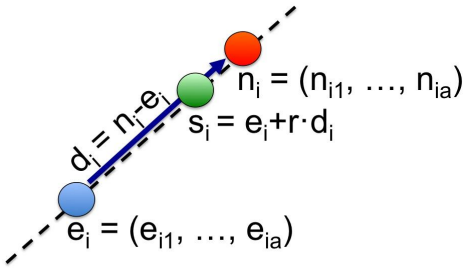


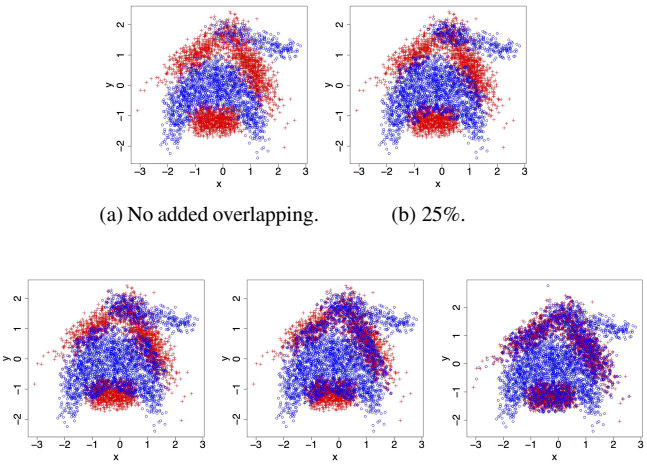
FIGURE 1: Creation of the synthetic sample s_i from two real samples e_i and n_i. The initial point e_i, a random number $r \in (0, 1)$ and the vector d_i between e_i and n_i are used.

introduction scheme proposed enables one to easily estimate which samples from O belong to the overlapping region and therefore to distinguish between overlapping and non-overlapping regions in the new dataset:

1) *Overlapping region* O_{ov}. The overlapping region is defined as the union of the synthetic samples S and the set B'_e composed of the original samples e_i used to create the synthetic samples (line 9 in Algorithm 1). Therefore, $O_{ov} = S \cup B'_e$.
2) *Non-overlapping region* $O_{\overline{ov}}$. It is formed by samples from D that are not included in B'_e, i.e., $O_{\overline{ov}} = D \setminus B'_e$.

Note that $O = O_{ov} \cup O_{\overline{ov}}$.

Figure 2 illustrates the result of applying the proposal to introduce overlapping in the `banana` dataset [9]. Several levels of overlapping have been introduced into one of its classes (the red one), from 0% (original data) to 100% (maximum overlapping), by increments of 25%. Note that the overlapping levels selected in this example play an illustrative role on the procedure of the proposal, although in real-world data the overlapping levels are not usually so high.



(a) No added overlapping. (b) 25%. (c) 50%. (d) 75%. (e) 100%.

FIGURE 2: Visualization of introducing different overlapping levels on the `banana` dataset.

C. OVERLAPPING INTRODUCTION SCHEMES

The proposed scheme can be applied to any class and also to any combination of classes in the data. This fact lead us to define different schemes to introduce overlapping, depending on the classes in which the scheme is applied. In this research two different overlapping schemes are considered:

1) *All-classes Overlapping Scheme* (AOS). This scheme individually considers each class of a dataset with C classes as the target class, resulting in C different sets of overlapping samples $S_1, \ldots, S_C$, which are finally merged with the original data D to build the overlapping dataset O ($O = D \cup S_1 \cup \ldots \cup S_C$). This scheme tries to simulate the most complex scenario in real-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2925300, IEEE Access

world problems, in which all the classes are overlapped with their surrounding classes –for example, as a result of a faulty sensor device that affects some attributes in all the samples in the dataset.

2) *Majority-class Overlapping Scheme* (MOS). In this scheme, overlapping is only introduced into the majority class, which is considered as the target class. The overlapping procedure results in a single set of synthetic samples S_{maj}, which is then added to the original dataset D to form the overlapping dataset O ($O = D \cup S_{maj}$). This scheme tries to model a situation in which the dataset has a difficult class, whose boundaries are not clearly defined. Therefore, the overlapping only affects to this class and its surrounding classes.

The procedure detailed in this section is used to generate new datasets with different levels and types of overlapping. All of them are then used to check the behavior of OVO when dealing with this type of data. However, this overlapping generation scheme can be used in future works to analyze any method trying to address the class overlapping problem. The design of the experimental framework and how the results obtained will be analyzed is described in Section V.

V. EXPERIMENTAL FRAMEWORK

First, Section V-A presents the base datasets considered for the experimentation, together with the types of overlapping and levels introduced into them. Then, Section V-B shows the algorithms used and their parameters. Finally, Section V-C explains the methodology for the analysis of the results.

A. REAL-WORLD DATASETS AND OVERLAPPING

The experimentation considers 34 real-world datasets from the UCI repository[1] [9], in which overlapping is introduced. Table 1 shows these datasets sorted by their number of classes (*cl*), along with the number of samples (*sa*) and attributes (*at*).

TABLE 1: Description of the real-world datasets used.

Dataset	*cl*	*sa*	*at*	Dataset	*cl*	*sa*	*at*
balance	3	625	4	dermatology	6	358	34
contraceptive	3	1473	9	flare	6	1066	11
hayesroth	3	160	4	glass	7	214	9
iris	3	150	4	satimage	7	6435	36
newthyroid	3	215	5	segment	7	2310	19
postoperative	3	87	8	zoo	7	101	16
splice	3	3190	60	ecoli	8	336	7
tae	3	151	5	marketing	9	6876	13
thyroid	3	7200	21	led7digit	10	500	7
wine	3	178	13	penbased	10	10992	16
car	4	1728	6	yeast	10	1484	8
lymphography	4	148	18	texture	11	5500	40
vehicle	4	846	18	vowel	11	990	13
cleveland	5	297	13	wqred	11	1599	11
nursery	5	12960	8	wqwhite	11	4898	11
pageblocks	5	5472	10	mlibras	15	360	90
automobile	6	159	25	abalone	28	4174	8

Motivated by the use of a *stratified k-fold cross validation* to estimate the classifier performance in the analysis of results, the creation of the k folds of the overlapping data O from the original data D is systematically carried out as follows:

1) A level of overlapping $x\%$ is used to create the set of synthetic samples S from the original dataset D following the scheme proposed in Section IV (either AOS or MOS). Overlapping levels from $x = 5\%$ to $x = 50\%$, by increments of 5% are considered.
2) The original data D and the set of synthetic samples S are partitioned with stratification into k folds each one, that is $D_1, \ldots, D_k$ and $S_1, \ldots, S_k$, respectively.
3) The k folds of the overlapping datasets are created as $O_i = D_i \cup S_i$, with $i = 1, \ldots, k$.

Note that the folds $D_1, \ldots, D_k$ in the original data D and those $O_1, \ldots, O_k$ in the overlapping dataset O have the same original samples in each fold $i = 1, \ldots, k$ for any overlapping type and level, being the synthetic samples of each fold the difference among them. In this way, a fairer comparison is established between different levels and types of overlapping over the same dataset, since possible differences in classifier performance due to a different partitioning in each level/type of overlapping are avoided as much as possible.

Once the k folds of the original dataset D and the overlapping dataset O have been obtained, two different ways of building the final dataset are considered depending on the folds affected by the overlapping (training or test partitions):

1) *Overlapping affecting training and test sets* (Section VI). The final overlapping dataset is formed by the folds $O_i (i = 1, \ldots, k)$. This situation represents the most realistic one in real-world data, in which both training and test sets may be affected by overlapping. This scenario allows one to check how classifiers behave in the different regions. Thus, in order to gain a deeper insight into the problem addressed, the analysis of these datasets is conducted in a three-step way: *(i)* by analyzing performance on all the samples (Section VI-A), *(ii)* on the samples from the non-overlapping regions (Section VI-B), and *(iii)* on the samples from the overlapping regions (Section VI-C). In this way we can check where is the contribution of OVO in terms of performance improvement.
2) *Overlapping affecting only the training sets* (Section VII). In this case, each test fold t is taken from the original data $D_t (t \in \{1, \ldots, k\})$, whereas the training set is composed of the remaining folds from the overlapping dataset. Introducing overlapping only into the training partitions while keeping the test partitions overlapping free allows one to observe how overlapping data affect the training process and how the test results are degraded depending on the type and level of overlapping (see Section VII-A). This scheme has also been used to deal with noisy data in classification [29].

As an outline, a total of 40 different configurations are applied to the 34 base datasets, resulting in a total of 1360 datasets with different types and levels of overlapping (1394

[1] http://archive.ics.uci.edu/ml/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2925300, IEEE Access

datasets if datasets without induced overlapping are also considered). Note that the overlapping level x = 0% is also studied, corresponding to the original datasets without additional induced overlapping. In detail, all the possible combinations among the following three factors are considered in the experiments:

1) *Sets affected by overlapping* (2): *(i)* training and test sets or *(ii)* only training sets.
2) *Overlapping schemes* (2): *(i)* AOS or *(ii)* MOS.
3) *Overlapping levels* (10): from $x = 5\%$ to $x = 50\%$, by increments of 5%.

B. CLASSIFICATION ALGORITHMS

Table 2 shows the classification algorithms considered for the experimentation along with their parameters setup, which is the recommended by their authors.

The choice of the learning algorithms has been made on the basis of their good behavior in a large number of real-world problems. They are classic and reference methods widely employed in many recent publications in the data mining literature [3], [5] and belong to different classification paradigms. C4.5 and RIPPER are rule-based classifiers, k−NN is a sample based learner and SVM builds hyperplanes to separate the transformed data in high-dimensional spaces. Note that the experiments performed are not focused on obtaining slightly better results by employing the most powerful algorithms, but checking whether OVO is able to improve the performance of the methods when data is affected by overlapping.

Two different values of k are used for the k−NN algorithm: $k = 3$ and $k = 5$. Notice that the value $k = 1$ is not considered in the experiments since 1−NN provides exactly the same classification results with or without the OVO decomposition. In this way, we can check how OVO is affected by different values of this important parameter when working with overlapping data.

TABLE 2: Classification algorithms and parameters.

Method	Ref.	Parameters
C4.5	[27]	confidence = 0.25, samples/leaf = 2, pruned tree
RIPPER	[6]	folds = 3, optimizations = 2
k−NN	[5]	neighbors = 3 and 5, distance = HVDM
SVM	[3]	polynomial kernel, C = 1, tol = 0.001, ϵ = 1.0E-12

C. METHODOLOGY OF ANALYSIS

In order to check the suitability of the methods using OVO when dealing with overlapping data, the results of the classification algorithms with and without decomposition are compared one another. For C4.5, RIPPER, 3−NN and 5−NN this comparison can be directly performed, since these techniques can handle multiple classes inherently. However, SVM is designed to work with binary datasets. For this reason, in the case of SVM, the OVO and OVA strategies are compared, checking which one of them has a better behavior with overlapping data.

The classifier accuracy estimation in each dataset is obtained running 5 iterations of a stratified 5-fold cross-validation (5x5-fcv). Each partition has a larger number of samples using 5 partitions and thus, the effects of overlapping samples become more notable. Furthermore, the 5 iterations of the 5-fcv enable the final results obtained to be as robust as possible. Hence, each overlapping dataset is created 5 times with different seeds, carrying out a total of 25 runs per dataset configuration, which are averaged to obtain the final result for each configuration and dataset. This fact implies that 34850 executions are carried out for each classifier (1394 datasets · 25 runs), which are repeated for the OVO and non-OVO versions, reaching a total of 348500 executions (5 classifiers, with OVO and non-OVO). For the sake of brevity, only averaged results are presented in this manuscript, but it must be stressed that the conclusions reached are based on the proper statistical analysis, which considers all the results (not averaged). Full results can be found on the web-page associated with this research [2].

The aforementioned analysis of the accuracy of each classifier is complemented by the study of the ELA [30] metric. This is a metric proposed in the framework of noisy data as a combination of the performance and robustness of the methods. It represents the robustness of the classifier when the noise level increases. This metric helps us check if a good performance is simultaneously combined with a good robustness, that is, the classifier is not so strongly deteriorated when higher levels of overlapping are considered. This metric is computed as

$$ELA_{x\%} = \frac{100 - Acc_{x\%}}{Acc_{0\%}}, \tag{3}$$

where $Acc_{x\%}$ is the accuracy with overlapping of $x\%$ and $Acc_{0\%}$ is the accuracy in the original data D. The ELA results are shown in percentage in this work, i.e. they are multiplied by 100.

In order to properly analyze the results obtained, *Wilcoxon*'s [8] non-parametric test is used. This is a pairwise test aiming at detecting significant differences between two sample means. For each one of the 40 configurations studied, the OVO and non-OVO versions are compared using *Wilcoxon*'s test and the p-values (p_W) associated are obtained. The p-value allows one to know whether two algorithms are significantly different. Even though the significance of the differences found is given by the p-value in *Wilcoxon*'s test, a threshold (significance level) is established to focus the analysis in the most interesting results. Thus, a difference will be considered to be significant if the p-value obtained is lower than 0.1, which is a value that usually shows important differences between the algorithms compared. Additionally, those cases in which the p-value is lower than 0.05 will also be analyzed, since these differences are even more meaningful than those at significance level 0.1.

[2] https://joseasaezm.github.io/overlapping/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2925300, IEEE Access

VI. ANALYSIS OF RESULTS OF OVERLAPPING DATA AFFECTING TRAINING AND TEST SETS

The experiments in this section deal with the scenario in which overlapping is introduced in both training and test sets. The main aim is to gain a full insight into the influence of overlapping on the classification process and the properties of the OVO decomposition mechanism in such a case. A three-stage analysis is designed to tackle the different aspects of this problem. Section VI-A analyzes the accuracy of classifiers on all the samples (those present in the original data and those overlapping samples synthetically generated). Section VI-B only focuses on the accuracy on those samples from the non-overlapping regions, whereas Section VI-C analyzes the accuracy on the samples from the overlapping regions. Finally, Section VI-D examines the robustness results of the classifiers considering the ELA measure.

A. OVERLAPPING IN TRAINING AND TEST SETS: ACCURACY ON ALL THE TYPES OF SAMPLES

This section studies the behavior of standard and OVO-based classifiers over all samples, i.e., those of the original data and those synthetically generated. Thus, the classification accuracy is measured in each dataset with respect to degree of overlapping between classes. Table 3 presents the accuracy of each method in its OVO and non-OVO version together with the output of *Wilcoxon*'s test.

The results in Table 3 show that, in general, applying the OVO decomposition leads to an improvement in accuracy over the standard single-model approach. A constant trend of the obtained accuracies is also observed, regardless of the overlapping level introduced. This fact shows the preferable characteristics of OVO-based learning, which can return an improved accuracy even in cases of extreme overlapping (50% of samples). This trend is further backed-up by *Wilcoxon*'s test, which shows that the gain from applying OVO is statistically significant for each type of base classifier, for almost all the overlapping levels in C4.5, RIPPER and SVM and for the highest overlapping levels in 3−NN and 5−NN (above 30% approximately). These results show the general stability of OVO when dealing with overlapping data.

Analyzing the cases with the AOS scheme, one can observe that the decrease in accuracy has similar trends for both original and OVO-based methods. However, OVO is always delivering an improved performance. AOS introduces artificial samples among all the classes leading to a complex scenario in which many classes may share similar sample distributions in given parts of the decision space. In such a situation, decomposition will lead to a significantly easier-to-solve case, where the classifier needs to deal with only two overlapping classes at once. However, even after such a simplification, the decision boundary is still not easy to estimate. Anyway, one must point out a strong potential advantage of this approach: it is significantly easier to perform data cleaning and transformation procedures on the overlapping region between only two classes. This allows one to conclude that OVO is a suitable approach for those cases when all the classes overlap, maintaining a better accuracy.

In the case of MOS, significantly higher accuracies for all the methods with respect to AOS are observed, as now the number of difficult regions is reduced. Generally, OVO is able to improve the performance of the methods –except for 3−NN, where OVO is statistically better only at the maximum overlapping level. The good performance of OVO in MOS can be explained by the fact that, after the decomposition, some class pairs will contain overlapping and some other will not. This fact allows for an improved classification performance, as standard multi-class classifiers can get their decision boundaries biased towards the overlapping cases. Notice that with OVO one may identify overlapping classes and apply data cleaning/transformation procedures only on the selected cases. This will reduce the complexity of the process in comparison to processing the whole multi-class dataset (in which some classes do not require any cleaning).

Note that the previous analysis is carried out considering both original and synthetic samples. In order to gain a deeper insight into the performance of OVO for overlapping data, next sections analyze whether the increased performance of OVO can be truly attributed to a greater robustness to overlapping or simply to a better classification of safe (non-overlapping) samples, as is traditionally checked in the literature.

B. OVERLAPPING IN TRAINING AND TEST SETS: ACCURACY IN NON-OVERLAPPING REGIONS

Table 4 presents the averaged accuracy results together with the output of *Wilcoxon*'s test when analyzing the non-overlapping (safe) original samples. In this case, the conclusions are drawn for both AOS and MOS simultaneously.

Analyzing the performance on samples from non-overlapping regions, one can observe that the accuracies of the classifiers are highly stable, as they work on safe data. The most notable exception to this fact is the behavior of RIPPER with the AOS scheme, which suffers from a high drop in performance if the OVO decomposition is not used. In this case, the increase in the predictive power of OVO can be purely attributed to its well-known properties of reducing the complexity of multi-class classification problems. As one can observe from both the obtained accuracies and *Wilcoxon*'s tests, OVO is able to boost the performance of all the classifiers. Only for certain cases with k−NN, differences are found to be statistically not significant. One must not forget that k−NN is a local classifier, since it only analyzes the neighborhood of a sample for its classification, and these methods do not work as well with OVO as global ones.

These results make mandatory the analysis of overlapping regions, as the performance improvement at this stage may be attributed only to the better behavior with safe samples.

C. OVERLAPPING IN TRAINING AND TEST SETS: ACCURACY IN OVERLAPPING REGIONS

In this section, we focus on analyzing the performance of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2925300, IEEE Access

TABLE 3: Accuracy results for all the samples, original and synthetic ones, with different types and ratios of overlapping in training and test sets –p_W are the p-values obtained, + denotes significant differences at level 0.1 and $*$ at level 0.05, and $<$ indicates that OVO obtains more ranks in *Wilcoxon*'s test and $>$ the opposite.

All-classes Overlapping Scheme

Method		0	5	10	15	20	25	30	35	40	45	50
ORI	C4.5	76.35	73.76	70.85	68.56	65.74	62.90	60.60	58.44	56.59	55.32	53.27
OVO		77.19	74.34	71.48	69.22	66.39	64.15	61.94	60.19	58.07	56.79	54.90
p_W		< 0.00216*	< 0.00933*	< 0.00492*	< 0.04256*	< 0.05232+	< 0.00030*	< 0.00005*	< 0.00001*	< 0.00005*	< 0.00023*	< 0.00013*
ORI	RIPPER	71.74	68.45	64.68	62.56	60.12	57.35	55.46	53.56	51.95	50.51	48.97
OVO		75.85	72.82	69.62	66.96	64.51	62.05	59.39	57.42	55.79	54.00	52.40
p_W		< 0.00001*	< 0.00001*	< 0.00000*	< 0.00000*	< 0.00000*	< 0.00000*	< 0.00000*	< 0.00000*	< 0.00000*	< 0.00000*	< 0.00000*
ORI	3–NN	74.77	71.61	67.87	64.76	61.27	57.70	55.23	52.79	50.16	48.32	45.78
OVO		76.28	73.09	69.15	66.19	63.04	59.36	56.86	54.13	51.59	49.72	47.19
p_W		< 0.17957	< 0.36942	< 0.80939	< 0.12181	< 0.02493*	< 0.03623*	< 0.00639*	< 0.05028+	< 0.01556*	< 0.09218+	< 0.01708*
ORI	5–NN	74.94	72.16	68.91	65.83	62.79	59.70	57.33	54.63	52.43	50.21	48.01
OVO		76.20	73.57	70.37	67.34	64.45	61.32	58.68	56.27	53.89	51.90	49.68
p_W		< 0.93868	< 0.73885	< 0.52145	< 0.71319	< 0.26279	< 0.25557	< 0.09553+	< 0.00285*	< 0.02148*	< 0.02803*	< 0.00397*
OVA	SVM	74.12	71.82	69.52	66.96	65.01	62.67	60.76	59.13	57.36	55.89	54.03
OVO		76.00	73.59	71.12	68.45	66.29	63.84	61.73	59.97	58.41	56.82	55.11
p_W		< 0.00030*	< 0.00088*	< 0.00113*	< 0.00746*	< 0.01349*	< 0.03329*	< 0.12221	< 0.14855	< 0.04105*	< 0.07681+	< 0.02683*

Majority-class Overlapping Scheme

Method		0	5	10	15	20	25	30	35	40	45	50
ORI	C4.5	76.35	75.17	73.88	72.64	71.76	70.52	69.15	68.47	67.55	66.60	65.63
OVO		77.19	75.86	74.55	73.29	72.43	71.22	70.01	69.41	68.60	67.50	66.96
p_W		< 0.00216*	< 0.00295*	< 0.00088*	< 0.00276*	< 0.01879*	< 0.00492*	< 0.00885*	< 0.00607*	< 0.00329*	< 0.01168*	< 0.00078*
ORI	RIPPER	71.74	70.20	68.55	67.20	66.41	64.96	63.90	62.98	62.49	61.54	60.68
OVO		75.85	74.36	72.73	71.65	70.62	69.60	68.20	67.30	66.32	65.50	64.31
p_W		< 0.00001*	< 0.00000*	< 0.00001*	< 0.00000*	< 0.00000*	< 0.00000*	< 0.00001*	< 0.00000*	< 0.00002*	< 0.00002*	< 0.00009*
ORI	3–NN	74.77	73.43	71.86	70.27	68.90	67.26	65.93	64.66	63.51	62.55	61.18
OVO		76.28	74.88	73.33	71.69	70.34	68.79	67.53	66.25	65.08	64.04	62.88
p_W		< 0.17957	< 0.54390	< 0.48868	< 0.40700	< 0.72598	< 0.25557	< 0.17957	< 0.39740	< 0.49948	< 0.19678	< 0.06120+
ORI	5–NN	74.94	73.71	72.23	70.90	69.46	68.29	66.80	65.60	64.43	63.33	62.27
OVO		76.20	74.96	73.59	72.26	70.91	69.68	68.34	67.27	66.15	65.21	64.11
p_W		< 0.93868	< 0.80421	< 0.44678	< 0.63824	< 0.36942	< 0.42662	< 0.04277*	< 0.09553+	< 0.04277*	< 0.01226*	< 0.01206*
OVA	SVM	74.12	72.82	71.76	70.34	69.45	68.28	67.31	66.56	65.70	64.86	63.85
OVO		76.00	74.74	73.51	72.19	71.09	69.91	68.82	67.98	67.23	66.37	65.35
p_W		< 0.00030*	< 0.00014*	< 0.00228*	< 0.00397*	< 0.00376*	< 0.00143*	< 0.00270*	< 0.00397*	< 0.00285*	< 0.01286*	< 0.00639*

OVO and non-OVO versions in the overlapping regions. Table 5 presents the averaged accuracy obtained with the output of *Wilcoxon*'s test. Note that for 0% degree of overlapping there are no samples to be analyzed.

Concentrating on the AOS scenario, a stable performance is observed for C4.5, RIPPER and SVM, regardless of the amount of overlapping introduced. This means that for both small and high degrees of overlapping these methods can return a similar fraction of correctly classified samples. This is a very desirable property as it proves the high robustness of these algorithms to the cases in which multiple classes overlap. The variation between the obtained results with increasing overlapping levels is always around 1%. For both k–NN classifiers the contrary behavior is observed. Their accuracies tend to significantly drop with increasing overlapping level, showing that these learners are not suitable for such a difficult scenarios.

When taking into account *Wilcoxon*'s test, OVO returns a statistically significant improvement over the original approach for C4.5 in almost all the overlapping levels; for RIPPER, 3–NN and 5–NN from 15-20% onwards; and for SVM in the maximum overlapping level. Additionally, a faster decrease in accuracies for both 3–NN and 5–NN with OVO than for their multi-class counterparts must be pointed out when overlapping increases (starting from around 3% of higher accuracy for OVO in 5% overlapping, and ending with only about 1-1.5% of gain for 50% of overlapping). This backs-up our previous claim that k–NN methods are not suitable for learning from overlapping datasets and that they do not work as well with OVO.

In the MOS scenario, only C4.5 is a stable learner, displaying identical characteristics as in the AOS case. 3–NN and 5–NN show identical correlation between the loss of their accuracies and the increase in the classification difficulty. The same behavior can be observed for SVM. It steadily loses accuracy, but in slower pace than NN-based approaches. This is an unexpected result, as it seems intuitive for the case with only some overlapping classes to be simpler than the AOS one. RIPPER displays a slightly higher variance than in the AOS scenario.

When analyzing the influence of OVO with *Wilcoxon*'s test, C4.5, 3–NN and 5–NN offer a significant improvement from an overlapping level of 25% onwards, whereas for RIPPER this fact occurs at some isolated levels. In this case, OVO

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2925300, IEEE Access

TABLE 4: Accuracy results for non-overlapping original samples, considering different types and ratios of overlapping in training and test sets –p_W are the p-values obtained, + denotes significant differences at level 0.1 and $*$ at level 0.05, and $<$ indicates that OVO obtains more ranks in *Wilcoxon*'s test and $>$ the opposite.

All-classes Overlapping Scheme

Method		0	5	10	15	20	25	30	35	40	45	50
ORI	C4.5	76.35	76.21	76.17	76.54	76.36	76.19	76.00	75.97	76.25	75.99	75.34
OVO		77.19	76.85	76.78	77.14	76.98	77.55	77.43	77.55	77.48	77.15	76.81
p_W		< 0.00216*	< 0.00840*	< 0.00492*	< 0.08269+	< 0.11378	< 0.00337*	< 0.00113*	< 0.00285*	< 0.04831*	< 0.06360+	< 0.00959*
ORI	RIPPER	71.74	71.14	70.11	70.05	69.76	68.95	69.35	68.70	68.24	67.82	67.98
OVO		75.85	75.24	75.11	75.03	75.13	75.08	74.76	74.40	74.60	74.16	74.37
p_W		< 0.00001*	< 0.00001*	< 0.00000*	< 0.00000*	< 0.00000*	< 0.00000*	< 0.00000*	< 0.00000*	< 0.00000*	< 0.00000*	< 0.00000*
ORI	3-NN	74.77	75.06	75.48	75.49	75.61	75.05	74.84	74.70	74.47	73.74	73.61
OVO		76.28	76.14	75.95	76.08	76.35	75.71	75.70	75.28	74.81	74.43	74.20
p_W		< 0.17957	< 0.53261	0.91152	< 0.30904	< 0.53261	< 0.13918	< 0.03623*	< 0.10253	< 0.07400+	< 0.19678	< 0.30098
ORI	5-NN	74.94	75.34	75.84	75.96	76.26	76.21	75.93	75.60	75.37	74.83	74.54
OVO		76.20	76.29	76.48	76.53	76.80	76.80	76.61	76.47	76.16	75.65	75.15
p_W		< 0.93868	0.84413	< 0.73885	< 0.76480	< 0.87098	< 0.67532	< 0.19678	< 0.07841+	< 0.19091	< 0.30904	< 0.28528
OVA	SVM	74.12	73.95	74.45	74.48	74.75	74.82	75.27	75.44	75.67	75.76	75.48
OVO		76.00	75.84	76.21	76.36	76.56	76.66	76.75	77.01	77.23	77.27	77.06
p_W		< 0.00030*	< 0.00044*	< 0.00094*	< 0.00171*	< 0.00356*	< 0.00397*	< 0.04831*	< 0.03056*	< 0.04105*	< 0.04634*	< 0.03056*

Majority-class Overlapping Scheme

Method		0	5	10	15	20	25	30	35	40	45	50
ORI	C4.5	76.35	76.29	76.17	76.11	76.25	76.15	75.84	76.10	76.17	76.05	75.95
OVO		77.19	77.01	76.88	76.79	76.86	76.57	76.43	76.89	77.04	76.71	76.89
p_W		< 0.00216*	< 0.00114*	< 0.00092*	< 0.00161*	< 0.01625*	< 0.06608+	< 0.03939*	< 0.03190*	< 0.01334*	< 0.06360+	< 0.01961*
ORI	RIPPER	71.74	71.38	70.83	70.63	70.67	70.43	70.13	69.99	70.31	70.02	70.05
OVO		75.85	75.52	75.14	75.37	75.50	75.23	75.06	75.01	75.14	74.98	74.90
p_W		< 0.00001*	< 0.00000*	< 0.00001*	< 0.00000*	< 0.00000*	< 0.00000*	< 0.00000*	< 0.00000*	< 0.00001*	< 0.00000*	< 0.00000*
ORI	3-NN	74.77	74.83	74.86	74.72	74.80	74.56	74.29	74.23	74.35	74.41	74.15
OVO		76.28	76.25	76.21	75.93	75.97	75.64	75.44	75.37	75.51	75.42	75.45
p_W		< 0.17957	< 0.58577	< 0.57846	< 0.59020	< 0.88446	> 0.89797	< 0.57846	> 0.99318	< 0.88446	< 0.77787	< 0.43663
ORI	5-NN	74.94	74.95	75.11	74.96	74.97	75.06	74.73	74.69	74.70	74.47	74.68
OVO		76.20	76.20	76.36	76.26	76.29	76.09	75.93	75.94	76.08	76.02	76.11
p_W		< 0.93868	< 0.84413	< 0.84413	< 0.81747	< 0.81747	< 0.92508	< 0.55530	< 0.48868	< 0.38793	< 0.16874	< 0.32558
OVA	SVM	74.12	73.78	73.87	73.57	73.73	73.59	73.60	73.76	73.99	73.98	74.05
OVO		76.00	75.74	75.72	75.58	75.62	75.46	75.45	75.56	75.73	75.86	76.00
p_W		< 0.00030*	< 0.00012*	< 0.00171*	< 0.00127*	< 0.00088*	< 0.00039*	< 0.00152*	< 0.00041*	< 0.00437*	< 0.00293*	< 0.00301*

generally fails to improve the classification process with SVM. Actually, OVA returns better accuracies on overlapping samples than OVO. This situation can be attributed to the fact that only partial overlapping between classes is present. In this way OVA always include the non-overlapping classes as counterexamples which may contribute towards better estimation of decision boundary for all the base classifiers (while OVO benefits from this only in some of its base learners).

D. OVERLAPPING IN TRAINING AND TEST SETS: ANALYSIS OF ROBUSTNESS OF THE CLASSIFIERS

This section analyzes the robustness of OVO to increasing overlapping levels. Table 6 presents the averaged ELA obtained, together with the output of *Wilcoxon*'s test (considering all the samples in the dataset).

When considering both the AOS and MOS scenarios, C4.5, RIPPER and SVM with OVO obtain a significantly better results than not using OVO. The stability of the differences between the standard and OVO versions should be remarked: OVO always performs better with almost the same difference for any level of overlapping. The differences between the improvement on accuracy and ELA should also be noticed. While on accuracy a gain of 2-4% is usually obtained using OVO, when applying ELA as metric, one observes up to 8% of gain in most of the cases. As ELA is designed for reflecting the performance of classifiers on noisy and difficult data, such a high gap proves the usefulness of applying OVO in scenarios where overlapping is to be expected in both training and testing sets. However, the situation is slightly different for 3−NN and 5−NN classifiers. Their OVO versions deliver a worse ELA performance and *Wilcoxon*'s test does not reject the null hypothesis of equivalence. OVO becomes significantly superior to its normal version only for some of the higher degrees of overlapping. This is another proof that minimal distance-based classifiers display lower robustness to overlapping and should not be used in such scenarios.

VII. ANALYSIS OF RESULTS OF OVERLAPPING DATA ONLY AFFECTING TRAINING SETS

This section assumes a scenario in which overlapping is introduced only in the training sets. This allows us to check how overlapping influences the learning process itself and how the estimated boundaries perform for normally distributed test samples. This way, we can examine the robustness of the training methods themselves and the importance of the training set quality. Section VII-A analyzes the accu-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2925300, IEEE Access

TABLE 5: Accuracy results for overlapping samples, considering different types and ratios of overlapping in training and test sets –p_W are the p-values obtained, + denotes significant differences at level 0.1 and $*$ at level 0.05, and $<$ indicates that OVO obtains more ranks in *Wilcoxon*'s test and $>$ the opposite.

All-classes Overlapping Scheme

Method		0	5	10	15	20	25	30	35	40	45	50
ORI	C4.5	✗	43.01	44.06	44.47	43.67	42.77	42.60	42.01	41.99	42.50	42.42
OVO		✗	42.34	44.40	45.14	44.43	43.83	43.87	43.93	43.64	44.23	44.11
p_W		✗	$>$ 0.75387	$<$ 0.08369+	$<$ 0.04153*	$<$ 0.01402*	$<$ 0.00028*	$<$ 0.00014*	$<$ 0.00000*	$<$ 0.00001*	$<$ 0.00008*	$<$ 0.00020*
ORI	RIPPER	✗	39.96	39.78	40.96	41.02	40.26	39.71	39.71	40.01	39.99	39.77
OVO		✗	42.07	42.07	42.72	42.64	42.41	41.58	41.54	41.79	41.64	41.59
p_W		✗	$<$ 0.29306	$<$ 0.51040	$<$ 0.09553+	$<$ 0.05887+	$<$ 0.00419*	$<$ 0.00057*	$<$ 0.00192*	$<$ 0.00519*	$<$ 0.00913*	$<$ 0.00419*
ORI	3-NN	✗	33.94	33.03	33.95	32.72	32.17	32.78	32.55	32.19	32.72	32.17
OVO		✗	36.56	35.97	36.71	35.67	34.62	34.93	34.30	34.11	34.38	33.83
p_W		✗	$<$ 0.12387	$<$ 0.42654	$<$ 0.01556*	$<$ 0.00073*	$<$ 0.03474*	$<$ 0.00106*	$<$ 0.02053*	$<$ 0.00746*	$<$ 0.01226*	$<$ 0.00152*
ORI	5-NN	✗	36.99	37.16	37.07	36.24	35.58	36.10	35.34	35.50	35.17	35.09
OVO		✗	40.16	40.07	39.79	39.06	37.97	37.74	37.42	37.24	37.19	37.09
p_W		✗	$<$ 0.00261*	$<$ 0.22809	$<$ 0.12598	$<$ 0.02683*	$<$ 0.01286*	$<$ 0.25557	$<$ 0.00065*	$<$ 0.04105*	$<$ 0.00466*	$<$ 0.00192*
OVA	SVM	✗	44.07	44.74	44.12	44.78	44.31	43.85	43.84	43.69	43.55	43.41
OVO		✗	44.74	45.50	44.57	44.99	44.58	44.22	43.95	44.32	44.09	44.18
p_W		✗	$<$ 0.35702	$<$ 0.13027	$<$ 0.37860	$<$ 0.41674	$<$ 0.93868	$<$ 0.84413	$<$ 0.67532	$<$ 0.20277	$<$ 0.07128+	$<$ 0.03939*

Majority-class Overlapping Scheme

Method		0	5	10	15	20	25	30	35	40	45	50
ORI	C4.5	✗	42.31	42.63	42.06	42.74	42.80	41.57	41.92	42.04	42.05	42.05
OVO		✗	41.84	42.30	42.28	43.92	45.00	44.03	43.57	43.82	43.52	44.22
p_W		✗	$>$ 0.66541	$<$ 0.82039	$<$ 0.97539	$<$ 0.18358	$<$ 0.00140*	$<$ 0.00183*	$<$ 0.01319*	$<$ 0.00140*	$<$ 0.00987*	$<$ 0.00013*
ORI	RIPPER	✗	38.84	40.07	38.79	40.54	39.42	39.87	40.28	40.59	40.67	40.20
OVO		✗	40.96	41.66	40.29	40.97	42.61	41.33	41.73	40.99	41.71	40.98
p_W		✗	$<$ 0.05770+	$<$ 0.39604	$<$ 0.29306	$>$ 0.66286	$<$ 0.03190*	$<$ 0.21516	$<$ 0.10253	$<$ 0.77787	$<$ 0.08269+	$<$ 0.35146
ORI	3-NN	✗	34.71	34.51	34.84	34.36	33.13	33.96	33.48	32.71	33.23	32.94
OVO		✗	35.06	35.91	36.39	35.98	35.39	35.92	35.40	34.40	34.98	34.50
p_W		✗	$<$ 0.77035	$<$ 0.68069	$<$ 0.12987	$<$ 0.78182	$<$ 0.04455*	$<$ 0.02349*	$<$ 0.01484*	$<$ 0.05942+	$<$ 0.00466*	$<$ 0.02734*
ORI	5-NN	✗	38.75	37.17	38.13	37.09	37.17	37.10	36.32	35.84	35.92	35.39
OVO		✗	39.73	39.08	39.10	38.57	39.29	39.03	38.35	37.81	38.10	37.49
p_W		✗	$<$ 0.34088	$<$ 0.05984+	$<$ 0.38970	$<$ 0.34821	$<$ 0.01415*	$<$ 0.02148*	$<$ 0.00959*	$<$ 0.01059*	$<$ 0.00285*	$<$ 0.00639*
OVA	SVM	✗	45.56	44.43	43.73	45.13	44.85	44.68	44.03	43.81	43.55	42.94
OVO		✗	44.75	44.33	44.23	44.76	44.62	44.28	43.48	44.05	43.45	42.71
p_W		✗	$<$ 0.95210	$>$ 0.97762	$>$ 0.92508	$>$ 0.36487	$>$ 0.5902	$>$ 0.19091	$>$ 0.16351	$<$ 0.44678	$>$ 0.407	$>$ 0.06120+

racy of classifiers on all the samples, whereas Section VII-B focuses on the ELA results.

A. OVERLAPPING IN TRAINING SETS: ANALYSIS OF ACCURACY ON ALL THE TYPES OF SAMPLES

Table 7 presents the averaged accuracy obtained, together with the output of *Wilcoxon*'s statistical test.

In AOS scenario, higher accuracies than the ones in Section VI are obtained. At the same time, increasing overlapping levels significantly influence the performance of the classifiers, however, once again not as much as in Section VI. This fact shows that overlapping only in the training set does not damage the classifier performance as strongly as the presence of this phenomenon in both sets, i.e., the real difficulty lies in predicting overlapped samples. Otherwise, the fact that in this case SVM is characterized by the smallest loss of accuracy should be highlighted.

SVM seems to be more robust to difficult training datasets than other classifiers, whereas this robustness is lost when difficult testing sets are faced (as shown in Section VI). In this case, the OVO approach offers a higher boost of accuracy, showing that by decomposing the multi-class dataset the training difficulties embedded within it may be alleviated. This makes OVO useful for working with uncertain input data. *Wilcoxon*'s test shows that the gains in accuracy when applying OVO are always statistically significant for SVM, C4.5 and RIPPER and at the highest overlapping levels for 3-NN and 5-NN.

The MOS scenario provides similar conclusions in all the cases but one, the effect of increasing overlapping level. A similar behavior of the classifiers for small overlapping ratios can be observed. However, the increase in the overlapping levels shows smaller drops in accuracy when compared to the AOS scenario. Note that testing samples are not affected by overlapping. This shows that OVO can efficiently deal with overlapping happening locally between only certain pairs of classes. Thus locally trained classifiers have a simplified task, since some of them will learn from safe cases without the presence of overlapping. Moreover, in this cases overlapping samples are less frequent, and therefore classifiers are less influenced by their presence at the same overlapping levels.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2925300, IEEE Access

TABLE 6: ELA robustness results (lower values are better) with all the samples and different types and ratios of overlapping in training and test sets –p_W are the p-values obtained, + denotes significant differences at level 0.1 and $*$ at level 0.05, and $<$ indicates that OVO obtains more ranks in *Wilcoxon*'s test and $>$ the opposite.

All-classes Overlapping Scheme

Method		0	5	10	15	20	25	30	35	40	45	50
ORI	C4.5	47.12	50.06	53.61	56.40	59.74	63.32	66.16	68.91	71.22	72.68	75.32
OVO		42.44	45.83	49.26	51.95	55.36	58.03	60.80	62.96	65.52	67.03	69.52
p_W		<0.00158*	< 0.00933*	< 0.00171*	<0.01971*	<0.01349*	< 0.00065*	< 0.00026*	< 0.00005*	<0.00030*	< 0.00032*	< 0.00036*
ORI	RIPPER	58.44	62.59	67.35	69.95	72.96	76.50	78.62	81.07	83.20	84.90	86.84
OVO		46.06	49.68	53.40	56.74	59.65	62.64	65.98	68.45	70.42	72.56	74.54
p_W		< 0.00001*	<0.00001*	< 0.00000*	< 0.00000*	< 0.00000*	< 0.00000*	< 0.00000*	< 0.00000*	<0.00000*	< 0.00000*	< 0.00000*
ORI	3-NN	55.26	58.96	63.49	66.99	70.94	75.48	78.22	81.08	83.98	86.27	89.63
OVO		48.51	52.25	56.97	60.42	63.94	68.70	71.54	74.91	77.82	80.11	83.29
p_W		> 0.22156	> 0.28528	> 0.67532	> 0.24155	< 0.07128+	> 0.11378	< 0.03190*	< 0.03778*	< 0.04455*	> 0.16351	< 0.08892+
ORI	5-NN	53.32	56.61	60.60	64.29	67.64	71.70	74.34	77.49	80.01	82.66	85.40
OVO		47.01	50.09	53.92	57.49	60.93	65.02	68.11	70.97	73.81	76.24	78.88
p_W		> 0.80421	> 0.72598	> 0.61402	> 0.75179	> 0.3974	> 0.34269	> 0.29306	< 0.06120+	> 0.16351	> 0.12181	> 0.12598
OVA	SVM	53.21	56.14	58.83	62.17	64.49	67.38	69.91	71.92	74.02	75.94	78.26
OVO		45.54	48.37	51.29	54.60	57.27	60.23	62.89	65.08	66.91	68.83	70.99
p_W		< 0.00015*	< 0.00050*	<0.00050*	< 0.00215*	<0.00241*	< 0.00492*	< 0.02247*	<0.01226*	< 0.00376*	< 0.01112*	< 0.00419*

Majority-class Overlapping Scheme

Method		0	5	10	15	20	25	30	35	40	45	50
ORI	C4.5	47.12	48.53	50.09	51.65	52.69	54.24	55.95	56.72	57.79	58.96	60.23
OVO		42.44	44.10	45.68	47.23	48.16	49.72	51.18	51.79	52.79	54.20	54.86
p_W		< 0.00158*	< 0.00191*	< 0.00073*	< 0.00134*	< 0.00463*	< 0.00356*	<0.00679*	< 0.00442*	< 0.00171*	< 0.00673*	< 0.00113*
ORI	RIPPER	58.44	60.24	62.25	63.90	64.96	66.68	68.04	69.24	69.67	70.89	71.90
OVO		46.06	47.92	49.82	51.08	52.26	53.56	55.27	56.40	57.52	58.50	59.91
p_W		< 0.00001*	< 0.00000*	< 0.00001*	< 0.00000*	<0.00000*	<0.00000*	< 0.00001*	< 0.00000*	< 0.00001*	<0.00001*	< 0.00001*
ORI	3-NN	55.26	56.85	58.59	60.42	62.02	64.16	65.64	67.17	68.31	69.48	71.14
OVO		48.51	50.22	51.99	53.90	55.60	57.43	59.02	60.46	61.74	63.15	64.42
p_W		> 0.22156	> 0.47801	> 0.5439	> 0.27764	> 0.5553	> 0.19678	> 0.24849	> 0.41674	> 0.48868	> 0.33407	<0.09898+
ORI	5-NN	53.32	54.80	56.59	58.08	59.87	61.24	63.03	64.50	65.65	67.19	68.24
OVO		47.01	48.54	50.11	51.67	53.28	54.76	56.47	57.74	58.97	60.22	61.42
p_W		> 0.80421	> 0.83078	> 0.77787	> 0.7648	> 0.61402	> 0.5553	> 0.23475	> 0.3974	> 0.13467	<0.08269+	<0.03939*
OVA	SVM	53.21	54.89	56.22	58.03	58.97	60.36	61.50	62.41	63.36	64.39	65.60
OVO		45.54	47.08	48.55	50.16	51.49	52.95	54.24	55.21	56.08	57.04	58.25
p_W		<0.00015*	< 0.00010*	<0.00044*	< 0.00088*	< 0.00078*	<0.00039*	< 0.00041*	< 0.00047*	<0.00028*	< 0.00215*	<0.00094*

B. OVERLAPPING IN TRAINING SETS: ANALYSIS OF ROBUSTNESS OF THE CLASSIFIERS

Table 8 presents the ELA results and the output of *Wilcoxon*'s test when analyzing the robustness of the OVO method with the ELA measure.

The analysis of ELA for both AOS and MOS shows the robustness of OVO to different overlapping levels in a similar way as in Section VI. In general, the overlapping affecting only to training sets has a strong impact on classifiers and OVO allows to generate a more robust set of base learners. C4.5, RIPPER and SVM combined with OVO return statistically significant improvements in comparison to their standard counterparts. In the case of RIPPER, the usage of OVO at the highest overlapping levels achieves an improvement of almost 15% of ELA. This fact shows the low robustness of RIPPER to difficult training sets, which can be easily improved using OVO. For 3-NN and 5-NN, OVO becomes better for overlapping levels greater than 30% in the most difficult case (AOS), showing greater differences than those observed in Section VI.

VIII. LESSONS LEARNED

This section summarizes the main findings on the usage of OVO [45], [46] from the empirical study in the previous sections:

1) **On the performance and robustness of OVO when dealing with overlapping data.** The methods that use OVO usually achieve higher performance results than their non-OVO counterparts, regardless of the overlapping level. The ELA metric [30] corroborates this conclusion showing greater differences than considering accuracy, as it takes the robustness of the method with respect to the case without overlapping into account. The robustness results are stable with respect to the overlapping level and strong variations on OVO performance are not observed. These facts show the suitability of OVO for overlapping scenarios.
2) **On the performance of OVO in the overlapping and non-overlapping regions.** The performance of the classifiers in these two regions shows that OVO is able to improve the accuracy on both sets. The overall performance improvement of OVO can be attributed to its well-known benefits in multi-class problems [13],

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2925300, IEEE Access

TABLE 7: Accuracy results for all the samples, original and synthetic ones, with different types and ratios of overlapping only in training sets –p_W are the p-values obtained, + denotes significant differences at level 0.1 and ∗ at level 0.05, and < indicates that OVO obtains more ranks in *Wilcoxon*'s test and > the opposite.

All-classes Overlapping Scheme

Method		0	5	10	15	20	25	30	35	40	45	50
ORI	C4.5	76.35	75.81	74.93	74.21	72.85	71.64	70.29	69.29	68.21	67.51	65.53
OVO		77.19	76.43	75.59	75.02	73.62	73.07	72.12	71.47	70.20	69.65	68.08
p_W		< 0.00216*	< 0.01089*	< 0.00840*	< 0.02803*	< 0.06360+	< 0.00161*	< 0.00020*	< 0.00008*	< 0.00044*	< 0.00135*	< 0.00030*
ORI	RIPPER	71.74	70.08	67.88	66.96	65.48	63.67	62.61	61.28	60.48	59.56	58.26
OVO		75.85	74.80	73.50	72.38	71.32	70.05	68.40	67.39	66.67	65.44	64.40
p_W		< 0.00001*	< 0.00000*	< 0.00000*	< 0.00000*	< 0.00000*	< 0.00000*	< 0.00000*	< 0.00000*	< 0.00000*	< 0.00000*	< 0.00000*
ORI	3-NN	74.77	73.60	71.76	70.08	67.52	64.94	63.04	60.70	58.40	56.34	53.38
OVO		76.28	75.17	73.11	71.61	69.63	66.94	65.01	62.54	60.49	58.57	55.70
p_W		< 0.17957	< 0.40699	< 0.76813	< 0.23475	< 0.04277*	< 0.03623*	< 0.02456*	< 0.05040+	< 0.03778*	< 0.01961*	< 0.02803*
ORI	5-NN	74.94	74.18	72.88	71.28	69.40	67.40	65.61	63.27	61.21	59.09	56.29
OVO		76.20	75.68	74.40	73.06	71.42	69.47	67.53	65.35	63.63	61.76	59.02
p_W		< 0.93868	< 0.71319	< 0.54390	< 0.20890	< 0.30904	< 0.18314	< 0.03329*	< 0.03474*	< 0.01059*	< 0.00607*	< 0.01349*
OVA	SVM	74.12	73.82	73.72	73.19	72.87	72.22	71.73	71.42	70.95	70.66	69.61
OVO		76.00	75.61	75.37	74.72	74.23	73.43	72.65	72.38	72.03	71.77	70.91
p_W		< 0.00030*	< 0.00083*	< 0.00135*	< 0.02349*	< 0.02927*	< 0.05028+	< 0.16874	< 0.16351	< 0.17409	< 0.15342	< 0.06120+

Majority-class Overlapping Scheme

Method		0	5	10	15	20	25	30	35	40	45	50
ORI	C4.5	76.35	76.12	75.68	75.24	74.89	74.25	73.47	73.14	72.39	72.15	71.38
OVO		77.19	76.84	76.42	76.03	75.54	75.01	74.39	74.37	73.76	73.10	73.09
p_W		< 0.00216*	< 0.00353*	< 0.00106*	< 0.00293*	< 0.02493*	< 0.02567*	< 0.01286*	< 0.00492*	< 0.00437*	< 0.05028+	< 0.00094*
ORI	RIPPER	71.74	70.78	69.71	68.91	68.32	67.37	66.72	66.13	65.84	65.38	64.87
OVO		75.85	75.20	74.26	73.89	73.30	72.73	72.05	71.39	70.95	70.35	69.92
p_W		< 0.00001*	< 0.00000*	< 0.00001*	< 0.00000*	< 0.00000*	< 0.00000*	< 0.00001*	< 0.00000*	< 0.00003*	< 0.00002*	< 0.00001*
ORI	3-NN	74.77	74.25	73.41	72.44	71.61	70.61	69.44	68.40	67.36	66.60	65.44
OVO		76.28	75.73	74.95	73.89	73.05	72.11	71.02	70.06	69.05	68.16	67.39
p_W		< 0.17957	< 0.57355	< 0.40606	< 0.53261	< 0.63824	< 0.33407	< 0.24155	< 0.30904	< 0.25557	< 0.13027	< 0.02148*
ORI	5-NN	74.94	74.57	73.90	73.16	72.29	71.66	70.39	69.53	68.30	67.37	66.67
OVO		76.20	75.87	75.28	74.64	73.78	73.05	71.95	71.20	70.18	69.45	68.64
p_W		< 0.93868	< 0.70682	< 0.47801	< 0.52145	< 0.22809	< 0.44678	< 0.10618	< 0.13467	< 0.05662+	< 0.01349*	< 0.02567*
OVA	SVM	74.12	73.86	73.85	73.20	73.01	72.50	72.21	71.98	71.51	71.15	70.41
OVO		76.00	75.80	75.56	75.08	74.55	73.98	73.55	73.28	72.85	72.36	71.65
p_W		< 0.00030*	< 0.00015*	< 0.00442*	< 0.00826*	< 0.03056*	< 0.02148*	< 0.06483+	< 0.08576+	< 0.06120+	< 0.16874	< 0.08576+

[28]. Moreover, focusing on overlapping samples, OVO alleviates the difficulties by considering classes by pairs, increasing their separability [12], [45].

3) **On the sets affected by overlapping (training/test).** Overlapping negatively affects the performance of the classifiers, independently of the sets where it is present (in training and test or only in training). However, classifiers have greater difficulties to deal with overlapping when both sets are affected, as correctly classifying test samples becomes harder. Likewise, overlapping in training affects the learning, producing more complex boundaries.
4) **On the amount of classes affected by overlapping.** Two different overlapping schemes have been studied with respect to the number of classes affected by overlapping: the AOS scheme (all classes are affected) and the MOS scheme (only the majority class is affected and consequently its surrounding classes.). As it could be expected, the AOS scheme has been generally more detrimental to classifier performance due to its higher complexity. In AOS, the usage of OVO allows us to reduce the multi-class problem to having only two overlapping classes in each base classifier, whereas in MOS some of the base classifiers are trained with non-overlapping pairs of classes. These facts contribute to the increase in performance of OVO.
5) **On the synergy between classifiers and OVO to deal with overlapping.** The behavior of five different classifiers (C4.5 [27], RIPPER [6], SVM [3], 3-NN and 5-NN [5]) has been studied with and without OVO in the presence of overlapping data. Three of them (C4.5, RIPPER and SVM) highly benefit from using OVO, providing good performance and robustness results on all levels of overlapping. More specifically, RIPPER obtains the highest improvements when OVO is used. 3-NN and 5-NN only benefit from OVO occasionally, but, in general, their performance is weaker than that of the other methods. Therefore, their usage should be avoided with overlapping data and they do not get the same advantage from decomposition strategies, mainly due to their local nature [5].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2925300, IEEE Access

TABLE 8: ELA robustness results (lower values are better) with all the samples and different types and ratios of overlapping in training sets –p_W are the p-values obtained, + denotes significant differences at level 0.1 and $*$ at level 0.05, and $<$ indicates that OVO obtains more ranks in *Wilcoxon*'s test and $>$ the opposite.

All-classes Overlapping Scheme

Method		0	5	10	15	20	25	30	35	40	45	50
ORI	C4.5	47.12	47.64	48.74	49.61	51.14	52.73	54.36	55.72	57.07	57.94	60.44
OVO		42.44	43.37	44.35	45.01	46.71	47.32	48.47	49.30	50.89	51.56	53.63
p_W		<0.00158*	< 0.00643*	< 0.00348*	<0.01226*	<0.04277*	<0.00113*	<0.00036*	<0.00013*	< 0.00061*	< 0.00094*	< 0.00032*
ORI	RIPPER	58.44	60.65	63.51	64.61	66.47	68.84	69.90	71.62	72.68	73.75	75.54
OVO		46.06	47.30	48.78	50.22	51.35	52.95	55.00	56.28	57.11	58.55	59.96
p_W		<0.00001*	< 0.00000*	< 0.00000*	< 0.00000*	<0.00000*	<0.00000*	<0.00000*	< 0.00000*	<0.00000*	< 0.00000*	< 0.00000*
ORI	3-NN	55.26	56.67	58.93	60.75	63.77	67.01	69.31	71.99	74.41	77.08	80.81
OVO		48.51	49.81	52.31	53.99	56.17	59.57	61.83	64.78	67.00	69.43	72.97
p_W		> 0.22156	> 0.30904	> 0.7648	> 0.30904	< 0.04455*	<0.06360+	<0.04831*	<0.04277*	<0.08892+	< 0.08269+	<0.10253
ORI	5-NN	53.32	54.28	55.99	57.95	60.13	62.72	64.84	67.59	69.91	72.46	75.92
OVO		47.01	47.60	49.20	50.74	52.67	55.23	57.48	60.11	61.98	64.30	67.54
p_W		> 0.80421	> 0.72598	> 0.5902	> 0.5104	> 0.33407	> 0.23475	<0.08892+	< 0.09553+	< 0.06360+	< 0.02456*	< 0.06864+
OVA	SVM	53.21	53.72	53.72	54.57	54.92	55.81	56.53	57.05	57.58	58.11	59.45
OVO		45.54	45.98	46.25	47.10	47.72	48.75	49.77	50.22	50.66	50.98	52.13
p_W		< 0.00015*	<0.00041*	< 0.00073*	< 0.00397*	<0.00826*	<0.01226*	<0.05662+	< 0.02148*	<0.02803*	<0.03623*	<0.01789*

Majority-class Overlapping Scheme

Method		0	5	10	15	20	25	30	35	40	45	50
ORI	C4.5	47.12	47.40	47.96	48.57	48.92	49.75	50.82	51.25	52.20	52.51	53.61
OVO		42.44	42.92	43.45	43.96	44.46	45.21	46.05	45.97	46.83	47.81	47.87
p_W		< 0.00158*	< 0.00203*	<0.00088*	< 0.00152*	<0.01089*	<0.01059*	<0.00913*	<0.00376*	< 0.00152*	<0.01630*	<0.00061*
ORI	RIPPER	58.44	59.58	60.98	62.00	62.88	64.03	64.96	65.72	66.02	66.69	67.42
OVO		46.06	46.91	48.03	48.39	49.06	49.90	50.73	51.61	52.12	52.95	53.47
p_W		<0.00001*	<0.00000*	<0.00000*	<0.00000*	<0.00000*	<0.00000*	<0.00001*	< 0.00000*	<0.00001*	< 0.00001*	<0.00001*
ORI	3-NN	55.26	55.91	56.83	58.02	58.92	60.33	61.70	63.03	64.11	65.10	66.60
OVO		48.51	49.23	50.13	51.40	52.49	53.65	55.08	56.18	57.30	58.64	59.47
p_W		0.22156	0.5104	0.5104	0.62608	0.5553	0.23475	0.33407	0.36942	0.30098	0.17957	<0.03474*
ORI	5-NN	53.32	53.82	54.70	55.58	56.66	57.46	59.12	60.17	61.57	62.94	63.59
OVO		47.01	47.45	48.16	48.94	49.98	50.93	52.35	53.26	54.50	55.57	56.47
p_W		> 0.80421	> 0.75179	> 0.70048	> 0.65051	> 0.36037	> 0.45706	> 0.25557	> 0.21516	> 0.12181	< 0.06864+	< 0.06864+
OVA	SVM	53.21	53.59	53.64	54.51	54.67	55.32	55.70	56.04	56.61	57.18	58.16
OVO		45.54	45.79	46.08	46.73	47.40	48.14	48.69	49.07	49.64	50.30	51.26
p_W		< 0.00015*	< 0.00012*	< 0.00083*	< 0.00152*	<0.00285*	< 0.00192*	<0.00492*	< 0.00492*	< 0.00547*	<0.02456*	< 0.01168*

IX. CONCLUDING REMARKS

In this research the problem of overlapping [16], [40] in the domain of multi-class classification [1], [42] is addressed. We suggest that using OVO [46] can improve the performance of base classifiers when treating problems with overlapping. In an exhaustive empirical study we have shown that OVO successfully helps in alleviating the influence of overlapping, without either needing to modify existing algorithms one by one or carrying out any prior data preprocessing step. Furthermore, to develop such an extensive study, we found the necessity for proposing a framework to introduce overlapping into real-world datasets. In this way, this framework is not only useful for the current study, but new developments in the field can also follow this new systematic way for creating a variety of classification problems with a measurable quantity of overlapping.

Our framework for introducing overlapping as well as our empirical study has considered two ways of introducing overlapping into existing datasets (in the majority class or in all the classes) and the possibility of adding it only in the training set or in both the training and test sets. All these combinations have allowed us to study the behavior of OVO in scenarios that display similar properties to real-world problems [33]. Decomposition performed by OVO helps to increase the separation between classes [13], [28] in these difficult to learn problems and it is beneficial to create more regular decision boundaries [45] where overlapping samples are present.

In future works we plan to develop data cleaning methods to reduce the difficulties in overlapping regions and sample weighting solutions to reduce the influence of overlapping samples on the decision boundaries given by the classifiers. We are specially interested in combining these approaches with decomposition strategies, where they can be applied to specific subproblems if needed.

ACKNOWLEDGMENT

José A. Sáez holds a *Juan de la Cierva-formación* fellowship (*Ref. FJCI-2015-25547*) from the Spanish Ministry of Economy, Industry and Competitiveness. Mikel Galar is partially supported by the Spanish Ministry of Economy, Industry and Competitiveness (MINECO), AEI/FEDER and UE under Project TIN2016-77356-P. Bartosz Krawczyk is partially supported by the Polish National Science Center under the

grant no. UMO-2015/19/B/ST6/01597.

REFERENCES

[1] F. Alswaina and K. Elleithy. Android malware permission-based multi-class classification using extremely randomized trees. IEEE Access, 6:76217–76227, 2018.

[2] C.K. Aridas, S.A.N. Alexandropoulos, S.B. Kotsiantis, and M.N. Vrahatis. Random resampling in the one-versus-all strategy for handling multi-class problems. Communications in Computer and Information Science, 744:111–121, 2017.

[3] M. Awad and R. Khanna. Support Vector Machines for Classification, pages 39–66. Apress, Berkeley, CA, 2015.

[4] G.E.A.P.A. Batista, R.C. Prati, and M.C. Monard. Balancing strategies and class overlapping. Lecture Notes in Computer Science, 3646 LNCS:24–35, 2005.

[5] G. Biau and L. Devroye. Lectures on the Nearest Neighbor Method. Springer Publishing Company, Inc., 2015.

[6] W.W. Cohen. Fast effective rule induction. In Proceedings of the 12th International Conference on Machine Learning, pages 115–123. Morgan Kaufmann Publishers, 1995.

[7] W.M. Czarnecki and J. Tabor. Two ellipsoid Support Vector Machines. Expert Systems with Applications, 41(18):8211–8224, 2014.

[8] J. Demšar. Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research, 7:1–30, 2006.

[9] D. Dua and C. Graff. UCI Machine Learning Repository (http://archive.ics.uci.edu/ml). University of California, Irvine, School of Information and Computer Sciences, 2019.

[10] A. Fernández, S. Garcia, F. Herrera, and N. Chawla. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. Journal of Artificial Intelligence Research, 61:863–905, 2018.

[11] M. Fu, Y. Tian, and F. Wu. Step-wise support vector machines for classification of overlapping samples. Neurocomputing, 155:159–166, 2015.

[12] J. Fürnkranz. Round Robin Classification. The Journal of Machine Learning Research, 2:721–747, 2002.

[13] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera. An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. Pattern Recognition, 44:1761–1776, 2011.

[14] V. García, R.A. Mollineda, J.S. Sánchez, R. Alejo, and J.M. Sotoca. When overlapping unexpectedly alters the class imbalance effects. Lecture Notes in Computer Science, 4478(2):499–506, 2007.

[15] V. García, J.S. Sánchez, and R.A. Mollineda. An empirical study of the behavior of classifiers on imbalanced and overlapped data sets. Lecture Notes in Computer Science, 4756:397–406, 2007.

[16] Y. Gu and L. Cheng. Classification of class overlapping datasets by kernel-mts method. International Journal of Innovative Computing, Information and Control, 13(5):1759–1768, 2017.

[17] X. Guan, J. Liang, Y. Qian, and J. Pang. A multi-view ova model based on decision tree for multi-classification tasks. Knowledge-Based Systems, 138:208–219, 2017.

[18] Y. Guo, H. Cao, S. Han, Y. Sun, and Y. Bai. Spectral-spatial hyperspectral image classification with k-nearest neighbor and guided filter. IEEE Access, 6:18582–18591, 2018.

[19] S. Huda, K. Liu, M. Abdelrazek, A. Ibrahim, S. Alyahya, H. Al-Dossari, and S. Ahmad. An ensemble oversampling model for class imbalance problem in software defect prediction. IEEE Access, 6:24184–24195, 2018.

[20] E. Hüllermeier and S. Vanderlooy. Combining predictions in pairwise classification: An optimal adaptive voting strategy and its relation to weighted voting. Pattern Recognition, 43(1):128–142, 2010.

[21] M. Kirk. Thoughtful Machine Learning. O'Reilly Media, Inc., Illinois, USA, 2014.

[22] R. Kretzschmar, N.B. Karayiannis, and F. Eggimann. Handling class overlap with variance-controlled neural networks. In International Joint Conference on Neural Networks, volume 1, pages 517–522, 2003.

[23] C.L. Liu. Partial discriminative training for classification of overlapping classes in document analysis. International Journal of Document Analysis and Recognition, 11(2):53–65, 2008.

[24] A. Lorena, A. de Carvalho, and J. Gama. A review on the combination of binary classifiers in multiclass problems. Artificial Intelligence Review, 30:19–37, 2008.

[25] R.C. Prati, G.E.A.P.A. Batista, and M.C. Monard. Class imbalances versus class overlapping: An analysis of a learning system behavior. In R. Monroy, G. Arroyo-Figueroa, L.E. Sucar, and H. Sossa, editors, MICAI 2004: Advances in Artificial Intelligence, pages 312–321, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[26] Y. Qu, H. Su, L. Guo, and J. Chu. A novel SVM modeling approach for highly imbalanced and overlapping classification. Intelligent Data Analysis, 15(3):319–341, 2011.

[27] J.R. Quinlan. C4.5: programs for machine learning. Morgan Kaufmann Publishers, San Francisco, CA, USA, 1993.

[28] R. Rifkin and A. Klautau. In defense of one-vs-all classification. Journal of Machine Learning Research, 5:101–141, 2004.

[29] J.A. Sáez, M. Galar, J. Luengo, and F. Herrera. Analyzing the presence of noise in multi-class problems: alleviating its influence with the One-vs-One decomposition. Knowledge and Information Systems, 38(1):179–206, 2014.

[30] J.A. Sáez, J. Luengo, and F. Herrera. Evaluating the classifier behavior with noisy data considering performance and robustness: The Equalized Loss of Accuracy measure. Neurocomputing, 176:26–35, 2016.

[31] B. Sidaoui and K. Sadouni. Binary tree multi-class SVM based on OVA approach and variable neighbourhood search algorithm. International Journal of Computer Applications in Technology, 55(3):183–190, 2017.

[32] W.Y. Sit, L.O. Mak, and G.W. Ng. Managing category proliferation in fuzzy artmap caused by overlapping classes. IEEE Transactions on Neural Networks, 20(8):1244–1253, 2009.

[33] J. Stefanowski. Dealing with Data Difficulty Factors While Learning from Imbalanced Data, pages 333–363. Springer International Publishing, Cham, 2016.

[34] W. Tang, K.Z. Mao, L.O. Mak, and G.W. Ng. Classification for overlapping classes using optimized overlapping region detection and soft decision. In 13th International Conference on Information Fusion, pages 1–8. IEEE Publishing, 2010.

[35] X. Tao, Q. Li, W. Guo, C. Ren, C. Li, R. Liu, and J. Zou. Self-adaptive cost weights-based support vector machine cost-sensitive ensemble for imbalanced data classification. Information Sciences, 487:31–56, 2019.

[36] P. Vorraboot, S. Rasmequan, K. Chinnasarn, and C. Lursinsap. Improving classification rate constrained to imbalanced data between overlapped and non-overlapped regions by hybrid algorithms. Neurocomputing, 152:429–443, 2015.

[37] D.R. Wilson and T.R. Martinez. Improved heterogeneous distance functions. Journal of Artificial Intelligence Research, 6(1):1–34, 1997.

[38] M. Woźniak, M. Grana, and E. Corchado. A survey of multiple classifier systems as hybrid systems. Information Fusion, 16(1):3–17, 2014.

[39] T.F Wu, C.J. Lin, and R.C. Weng. Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research, 5:975–1005, 2004.

[40] H. Xiong, M. Li, T. Jiang, and S. Zhao. Classification algorithm based on NB for class overlapping problem. Applied Mathematics and Information Sciences, 7(2 L):409–415, 2013.

[41] Y. Yan, R. Liu, Z. Ding, X. Du, J. Chen, and Y. Zhang. A parameter-free cleaning method for SMOTE in imbalanced classification. IEEE Access, 7:23537–23548, 2019.

[42] S. Yang, C. Zhang, and W. Wu. Binary output layer of feedforward neural networks for solving multi-class classification problems. IEEE Access, 7:5085–5094, 2019.

[43] Z. Yang and D. Gao. Classification for imbalanced and overlapping classes using outlier detection and sampling techniques. Applied Mathematics and Information Sciences, 7(1 L):375–381, 2013.

[44] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang. Efficient knn classification with different numbers of nearest neighbors. IEEE Transactions on Neural Networks and Learning Systems, 29(5):1774–1785, 2018.

[45] Z.L. Zhang, X.G. Luo, S. García, J.F. Tang, and F. Herrera. Exploring the effectiveness of dynamic ensemble selection in the one-versus-one scheme. Knowledge-Based Systems, 125:53–63, 2017.

[46] L. Zhou, Q. Wang, and H. Fujita. One versus one multi-class classification fusion using optimizing decision directed acyclic graph for predicting listing status of companies. Information Fusion, 36:80–89, 2017.